# TITLE

- **S&T TEXT MINING**

- DR. RONALD N. KOSTOFF

- OFFICE OF NAVAL RESEARCH

- PRESENTATION TO STIC

- 11 JANUARY 2001

# OUTLINE

- **DEFINITIONS/ GOALS**
- CAPABILITIES/ EXAMPLES
- CROSSOVER SCIENCE
  - BACKGROUND
  - CONCEPT
  - PROPOSAL
- DEFICIENCIES
- NEXT STEPS
- SUMMARY

# DEFINITIONS/ GOALS

- TM DEFINITIONS
- DATA MINING: EXTRACTION OF USEFUL INFORMATION FROM DATA
- TEXT MINING: EXTRACTION OF USEFUL INFORMATION FROM TEXT
  - COMPUTER-BASED, LARGE VOLUMES
- S&T TEXT MINING: EXTRACTION OF USEFUL INFORMATION FROM TECHNICAL TEXT
- ADDED COMPLEXITY: NEED FOR LEXICON, CONTEXT

# DEFINITIONS/ GOALS

- TM COMPONENTS
  - INFORMATION RETRIEVAL
  - INFORMATION PROCESSING
    - BIBLIOMETRICS
    - COMPUTATIONAL LINGUISTICS
    - CLUSTERING
  - INFORMATION INTEGRATION

# DEFINITIONS/ GOALS

- TWO APPROACHES
  - SOCIOLOGICAL
    - HIGH LEVEL OVERVIEW
    - LOW RESOLUTION RESULTS
    - HIGH FREQUENCY PHENOMENA
    - MODEST INPUTS OF TECHNICAL EXPERTISE
    - AMENABLE TO SEMI-AUTOMATED ANALYSIS
    - SHORT TIME REQUIRED
    - RELATIVELY LOW COST
    - LITTLE NEW INFORMATION TO TECHNICAL EXPERTS

# DEFINITIONS/ GOALS

- ANALYTICAL
  - DETAILED INSIGHTS
  - HIGH RESOLUTION RESULTS
  - LOW FREQUENCY PHENOMENA
  - SUBSTANTIAL INPUTS OF TECHNICAL EXPERTISE
  - MORE MANUAL EFFORTS REQUIRED
  - LONGER TIME REQUIRED
  - MODEST COST REQUIRED
  - NEW INFORMATION AND INSIGHTS FOR TECHNICAL EXPERT

# DEFINITIONS/ GOALS

FULL ACCESS AND INSIGHT TO RELEVANT GLOBAL S&T DATA TO SUPPORT:

1) DISCOVERING AND INNOVATING FROM LITERATURE,

2) PLANNING/ EXECUTING/ MANAGING/ TRANSITIONING OF S&T

# DEFINITIONS/ GOALS

- **HELP ANSWER FOLLOWING GENERIC QUESTIONS:**
  - WHAT S&T IS BEING DONE GLOBALLY?
  - WHO IS DOING IT?
  - WHERE IS IT BEING DONE?
  - WHAT MESSAGES CAN BE EXTRACTED FROM GLOBAL S&T?
  - WHAT PROMISING DIRECTIONS CAN BE IDENTIFIED?
  - WHAT IS NOT BEING DONE?
  - **--->WHAT SHOULD WE BE DOING DIFFERENTLY?**

# DEFINITIONS/ GOALS

- **RETRIEVE S&T DOCUMENTS FROM GLOBAL DATABASES**
  - SCI, COMPENDEX, WEB, NTIS, RADIUS, MEDLINE
- **IDENTIFY TECHNOLOGY INFRASTRUCTURE**
  - AUTHORS, JOURNALS, ORGANIZATIONS, ETC
  - REVIEW PANELS, WORKSHOPS, SITE VISITS
- **IDENTIFY CITATION NETWORKS**
  - IMPACT TRACKING, SPONSOR PRESENTATIONS
- **LITERATURE-BASED DISCOVERY**
  - PROMISING S&T DIRECTIONS/ OPPORTUNITIES
- **IDENTIFY PERVASIVE SUB-TECHNOLOGY THEMES**
  - ESTIMATE RELATIVE GLOBAL LEVELS OF EMPHASIS
  - GENERATE TAXONOMIES
- **IDENTIFY THEME RELATIONSHIPS**
  - CLUSTERING OF COMMON THEMES
  - GENERATE BOTTOM-UP TAXONOMIES
  - ALSO INTEL APPLICATIONS
  - SUPPORTS PROGRAM/ ORGANIZATIONAL RE-STRUCTURING

# OUTLINE

- DEFINITIONS/ GOALS
- **CAPABILITIES/ EXAMPLES**
- CROSSOVER SCIENCE
  - BACKGROUND
  - CONCEPT
  - PROPOSAL
- DEFICIENCIES
- NEXT STEPS
- SUMMARY

# CAPABILITIES/ EXAMPLES

- **INFORMATION RETRIEVAL - PRODUCT**
  - COMPREHENSIVE RECORDS
  - HIGHLY RELEVANT RECORDS
  - MULTIPLE DATABASES
    - SCI
    - EC
    - NTIS
    - MEDLINE
  - COMPLETE QUERY

# CAPABILITIES/ EXAMPLES

- **INFORMATION RETRIEVAL - PROCESS**
  - START WITH INITIAL TEST QUERY
  - EXPERT DIVIDES RECORDS RETRIEVED INTO RELEVANT/ NON-RELEVANT
  - OBTAIN PATTERNS CHARACTERISTIC OF EACH GROUP (LINGUISTIC/ BIBLIOMETRIC)
    - RELEVANT GROUP PATTERNS PROVIDE COMPREHENSIVENESS
    - NON-RELEVANT GROUP PATTERNS ELIMINATE NOISE RECORDS
  - ITERATE UNTIL CONVERGENCE OBTAINED
  - ***MOST CRITICAL PART OF TEXT MINING***

# CAPABILITIES/ EXAMPLES

- **INFORMATION RETRIEVAL - EXAMPLE**
- SHIP HYDRODYNAMICS
- (hydrodynamic* or hydromechanic* or fluid flow or potential flow or incompressible flow or wake or turbulen* or vort*) AND (bound* or ship* or surface* or hull* or fish or dolphin) NOT (accret* or adhes* or adsor* or aggregat* or bacter* or bear* or black hole or carbon* or cluster* or colli* or colloid* or combustion or crystal* or dissol* or emiss* or erosion or flame* or fractur* or gala* or grain* or ion* or larva* or lubrica* or melt* or membrane* or microscop* or mineral* or molecul* or organ* or permea* or plasm* or poro* or protein* or rock* or sediment* or shell* or shock or star or stars or stellar or sulf* or surface brightness or weld* or x-ray ageostrophic or animal* or antarctic or arctic or bay or bio* or cancer or CFC* or cilia or climat* or cloud* or coloni*or cosm* or crack* or cultivation or cumulus or diatom* or DNA or dunes or earthquake* or eco* or fermi or fluidised bed* fluidized bed* or greenhouse or gyre* or hydrographic or intertidal or Josephson or leaf or liposome* or monsoon* or muddy or nucl* or nutrient* or ozone or photolysis or phytoplankton or quantum or Rossby or sand or snow or soil or strato* or superconduct* or tropopause or undercurrent or ventricular or volcan* or zoo* or ablation or agglomeration or algal or alto* or astro-physics or astronomy or Benard convection or baroclinic* or barotropic* or blood flow or botan* or Brownian motion or capillary or cardiolog* or carotid or casting or CCD or cells or computational combustion dynamics or condensation or cyclon* or Darcy* or deep drawing or deposition or drainage or dredg* or drying or Ekman or electrochem* or environment*or enzyme* or estuary flow or fault* or film or foundry or fractal* or geostrophic or glycolipid* or granular or groundwater or Gulf-stream or heart or hydrology or hypersonic or ice mechanics or insect or irrigation or Kelvin-Helmholtz or laser welding or lipid* or liquid metal* or liquid-metal or locomotion or mantle or manufact* or materials or medical or microgravity or micromolecular or microscale or mining or molding or molten or Oseen or osmosis or physiolog* or pollution or polyphase flow or powder or preditor* or protozoa or pylori* or rain* or rarefied gas or reacting flow* or refuse or resuspension or roller* or rolling or scour* or seals or seismic or siltation or sintering or slag or solar or soldering or solenoid* or solidification or storm or sun or superfluid or supersonic or suspension* or tecton* or tide* or tidal or tokamak or tribology or turbidity or ultrasonic* or upwelling)

# CAPABILITIES/ EXAMPLES

- **INFORMATION RETRIEVAL - EXAMPLE**
- AIRCRAFT
  - SCIENCE CITATION INDEX
    - APPROXIMATELY 5600 JOURNALS & MAGAZINES.
    - PHYSICAL, ENGINEERING & LIFE SCIENCES BASIC RESEARCH.
    - 1991 - MID 1998.
    - PRODUCED 4346 APPLICABLE RECORDS
  - .ENGINEERING COMPENDEX
    - APPROXIMATELY 2600 JOURNALS & CONFERENCE PROCEEDINGS.
    - MAINLY APPLIED RESEARCH AND TECHNOLOGY.
    - 1990 - MID 1998
    - PRODUCED 15,673 APPLICABLE RECORDS.

# CAPABILITIES/ EXAMPLES

- **INFORMATION RETRIEVAL - EXAMPLE**
- AIRCRAFT (CONT'D)
  - SCI
    - REQUIRED SIGNIFICANT EFFORT TO DEVELOP QUERY FOR COMPREHENSIVE HIGH S/N RELEVANT RECORDS
    - REQUIRED A QUERY THAT CONSISTED OF 207 TERMS
    - STARTED WITH "AIRCRAFT" ; SUBTRACTED NON-RELEVANT TERMS
  - EC
    - CONSIDERABLY MORE FOCUSED ON JOURNALS/ PUBLICATIONS OF INTEREST. VERY FEW EXTRANEOUS RECORDS GENERATED WITH 13 TERM QUERY.

- ***COMPLEXITY OF QUERY DEPENDS ON RELATION OF DATABASE CONTENTS TO OBJECTIVES OF STUDY.***

# CAPABILITIES/ EXAMPLES

- **INFORMATION RETRIEVAL - BENEFITS**
- ITERATIVE QUERY APPROACH ALLOWS:
  - INCREASED RATIO OF RELEVANT/ NON-RELEVANT RECORDS; HIGHER SIGNAL-TO-NOISE RATIO
    - NOISE REDUCTION VERY IMPORTANT FOR LARGE RETRIEVALS
    - IMPROVES ANALYSIS RESULTS - KET LAW
  - MORE RECORDS IN FOCUSED FIELD TO BE RETRIEVED; INCREASED SIGNAL
    - USES LANGUAGE OF AUTHORS
  - MORE RECORDS IN ALLIED FIELDS TO BE RETRIEVED
  - POTENTIALLY RELEVANT RECORDS IN DISPARATE FIELDS TO BE RETRIEVED

# CAPABILITIES/ EXAMPLES

- **BIBLIOMETRICS - PRODUCT**
  - PROLIFIC AUTHORS
  - JOURNALS CONTAINING RELEVANT PAPERS
  - ORGANIZATIONS PRODUCING RELEVANT PAPERS
  - COUNTRIES PRODUCING RELEVANT PAPERS
  - MOST CITED AUTHORS
  - MOST CITED PAPERS
  - MOST CITED JOURNALS

# CAPABILITIES/ EXAMPLES

- **BIBLIOMETRICS - PROCESS**
  - START WITH RETRIEVED RECORDS
  - COMPUTE OCCURRENCE FREQUENCIES
  - GENERATE LISTS
  - GENERATE DISTRIBUTION FUNCTIONS
  - COMPARE WITH OTHER STUDIES

# CAPABILITIES/ EXAMPLES

- **BIBLIOMETRICS - EXAMPLES**
  - MOST CITED AUTHORS - AIRCRAFT
    **(**CITED BY OTHER PAPERS IN DATABASE**)**
    - ERICSSON-LE,117
    - JOHNSON-W,97
    - MIELE-A,96
    - DOYLE-JC,82
    - TISCHLER-MB,80
    - SRINIVASAN-GR,78
    - PETERS-DA,75
    - HODGES-DH,70
    - HESS-RA,60
    - FRIEDMANN-PP,55
    - CHATTOPADHYAY-A,55
    - NEWMAN-JC,54
    - FARASSAT-F,53
    - JAMESON-A,50
    - MENON-PKA,50

# CAPABILITIES/ EXAMPLES

- BIBLIOMETRICS - EXAMPLES
  - MOST CITED AUTHORS - FULLERENES
    - KROTO HW,4328
    - KRATSCHMER W,3472
    - IIJIMA S,1787
    - TAYLOR R,1721
    - HADDON RC,1711
    - HEBARD AF,1563
    - DIEDERICH F,1476
    - FOWLER PW,1469
    - BETHUNE DS,1466
    - HIRSCH A,1264
    - EBBESEN TW,1145
    - ALLEMAND PM,1103
    - HEINEY PA,1064
    - HAUFLER RE,1021

# CAPABILITIES/ EXAMPLES

- **BIBLIOMETRICS - EXAMPLES**
  - MOST CITED PAPERS - AIRCRAFT
    - 'JOHNSON-W,1980,HELICOPTER-THEORY',28
    - 'SNELL-SA,1992,J-GUID-CONTROL-DYNAM,V15',25
    - 'DOYLE-JC,1989,IEEE-T-AUTOMAT-CONTR,V34',23
    - 'LANE-SH,1988,AUTOMATICA,V24',22
    - 'ISIDORI-A,1989,NONLINEAR-CONTROL-SY',20
    - 'MCRUER-D,1973,AIRCRAFT-DYNAMICS-AU',19
    - 'KWAKERNAAK-H,1972,LINEAR-OPTIMAL-CONTR',18
    - 'DOYLE-JC,1981,IEEE-T-AUTOMAT-CONTR,V26',18
    - 'MACIEJOWSKI-JM,1989,MULTIVARIABLE-FEEDBA',17
    - 'MEYER-G,1984,AUTOMATICA,V20',17
    - 'GOLDBERG-DE,1989,GENETIC-ALGORITHMS-S',17
    - 'BRYSON-AE,1975,APPLIED-OPTIMAL-CONT',17
    - 'MENON-PKA,1987,J-GUID-CONTROL-DYNAM,V10',16
    - 'MCLEAN-D,1990,AUTOMATIC-FLIGHT-CON',16
    - 'NARENDRA-KS,1990,IEEE-T-NEURAL-NETWOR,V1',16
    - 'VANDERPLAATS-GN,1984,NUMERICAL-OPTIMIZATI',15

# CAPABILITIES/ EXAMPLES

- **BIBLIOMETRICS - EXAMPLES**
  - MOST CITED PAPERS - FULLERENES
    - KRATSCHMER W 1990 NATURE V347,2773
    - KROTO HW 1985 NATURE V318,2319
    - HEBARD AF 1991 NATURE V350,1177
    - IIJIMA S 1991 NATURE V354,816
    - HEINEY PA 1991 PHYS REV LETT V66,742
    - HAUFLER RE 1990 J PHYS CHEM US V94,720
    - ALLEMAND PM 1991 J AM CHEM SOC V113,683
    - AJIE H 1990 J PHYS CHEM US V94,659
    - HADDON RC 1991 NATURE V350,602
    - KRATSCHMER W 1990 CHEM PHYS LETT V170,556
    - SAITO S 1991 PHYS REV LETT V66,527
    - KROTO HW 1991 CHEM REV V91,507
    - FLEMING RM 1991 NATURE V352,504

# CAPABILITIES/ EXAMPLES

- **BIBLIOMETRICS - BENEFITS**
  - CRITICAL INFRASTRUCTURE IDENTIFIED
    - SELECTION OF CREDIBLE EXPERTS FOR WORKSHOPS/ REVIEW PANELS
    - IDENTIFICATION OF PRODUCTIVE PEOPLE AND ORGANIZATIONS FOR SITE VISITS
  - PRODUCTIVITY AND IMPACT TRACKING
  - INTELLECTUAL HERITAGE IDENTIFICATION

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - PRODUCT**
  - PERVASIVE TECHNICAL THEMES
  - RELATIONS AMONG THEMES
  - RELATIONS AMONG TECHNICAL THEMES AND INFRASTRUCTURE
  - TAXONOMIES
  - GLOBAL LEVELS OF EMPHASIS

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - PROCESS**
- PERVASIVE TECHNICAL THEMES
  - PHRASE FREQUENCY ANALYSIS
  - SELECT HIGH TECHNICAL CONTENT PHRASES
  - SELECT HIGH FREQUENCY PHRASES

# CAPABILITIES/ EXAMPLES
# PERVASIVE TECHNICAL THEMES
## AIRCRAFT S&T

### One Word

| | |
|---|---|
| 1178 | AIRCRAFT |
| 554 | CONTROL |
| 253 | PERFORMANCE |
| 219 | HELICOPTER |
| 198 | ROTOR |
| 178 | COMPOSITE |
| 176 | STRUCTURES |
| 154 | ENGINE |
| 149 | MATERIALS |
| 149 | RESPONSE |
| 146 | TEST |
| 143 | SIMULATION |
| 142 | DAMAGE |
| 140 | STRUCTURAL |
| 137 | TECHNOLOGY |
| 133 | DYNAMICS |
| 127 | NOISE |
| 123 | DYNAMIC |
| 123 | NONLINEAR |
| 119 | AERODYNAMIC |

### Two Word

| | |
|---|---|
| 71 | FLIGHT CONTROL |
| 65 | FINITE ELEMENT |
| 60 | CONTROL SYSTEM |
| 40 | GAS TURBINE |
| 38 | AIRCRAFT STRUCTURES |
| 38 | CONTROL SYSTEMS |
| 38 | HELICOPTER ROTOR |
| 37 | NEURAL NETWORK |
| 35 | HANDLING QUALITIES |
| 30 | EXPERIMENTAL DATA |
| 29 | CRACK GROWTH |
| 29 | TRANSPORT AIRCRAFT |
| 27 | BOUNDARY LAYER |
| 27 | NEURAL NETWORKS |
| 26 | FLIGHT TEST |
| 25 | AIRCRAFT ENGINES |
| 25 | AIRCRAFT GAS |
| 25 | FATIGUE DAMAGE |
| 25 | FIGHTER AIRCRAFT |
| 25 | FRACTURE MECHANICS |

### Three Word

| | |
|---|---|
| 29 | FLIGHT CONTROL SYSTEM |
| 19 | AIRCRAFT GAS TURBINE |
| 15 | THERMAL BARRIER COATINGS |
| 14 | COMPUTATIONAL FLUID DYNAMICS |
| 14 | FINITE ELEMENT METHOD |
| 13 | FLIGHT CONTROL SYSTEMS |
| 13 | QUANTITATIVE FEEDBACK THEORY |
| 12 | ANGLE OF ATTACK |
| 12 | ELEMENT ALTERNATING METHOD |
| 12 | FINITE ELEMENT ALTERNATING |
| 12 | HOVER AND FORWARD |
| 11 | EQUATIONS OF MOTION |
| 11 | FATIGUE CRACK GROWTH |
| 11 | GAS TURBINE ENGINES |
| 10 | ELASTIC-PLASTIC FINITE ELEMENT |
| 10 | FLIGHT TEST DATA |
| 10 | GAS TURBINE ENGINE |
| 10 | MICROSTRUCTURE AND PROCESSING |
| 10 | MULTIPLE SITE DAMAGE |
| 10 | WIDESPREAD FATIGUE DAMAGE |

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - PROCESS**
- RELATIONS AMONG THEMES
  - SELECT PHRASES OF PARTICULAR INTEREST (THEMES) FROM PHRASE FREQUENCY ANALYSIS, BASED ON STUDY OBJECTIVES
  - IDENTIFY PHRASES LOCATED PHYSICALLY CLOSE TO THE THEME PHRASES THROUGHOUT THE TEXT
  - USE NUMERICAL INDICATORS TO FILTER OUT THOSE PHRASES MOST CLOSELY ASSOCIATED WITH THEME PHRASE
  - PROVIDES ESTIMATES OF STRENGTH OF ASSOCIATION OF TEXT PHRASES TO THEME PHRASE

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - EXAMPLE (NEAR-EARTH SPACE STUDY)**
- RELATION AMONG THEMES (***REMOTE SENSING***)
  - APPLICATIONS (DETECTION OF OIL SLICKS, MONITORING FREEZE-THAW CYCLES, VEGETATION MAPPING)
  - REGIONS (COASTAL ENVIRONMENTS, TROLLFJORD-KOMAGLEV FAULT ZONE, VARANGER PENINSULA, AURORAL ZONES, TERRESTRIAL ECOSYSTEMS)
  - FEATURES (SURFACE MINING, WHEAT ACREAGE, DARK DENSE VEGETATION, SNOW HYDROLOGY, COAL MINING, CORAL REEF, UNSTRESSED CANOPY, BLACK SPRUCE PICEA MARIANA)

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - PROCESS**
- RELATIONS AMONG TECHNICAL THEMES AND INFRASTRUCTURE
  - SELECT PHRASES OF PARTICULAR INTEREST (THEMES) FROM PHRASE FREQUENCY ANALYSIS, BASED ON STUDY OBJECTIVES
  - IDENTIFY INFRASTRUCTURE TERMS LOCATED PHYSICALLY CLOSE TO THE THEME PHRASES THROUGHOUT THE DATABASE OF ***NON-ABSTRACT*** FIELDS
  - USE NUMERICAL INDICATORS TO FILTER OUT THOSE INFRASTRUCTURE TERMS MOST CLOSELY ASSOCIATED WITH THEME PHRASE
  - PROVIDES ESTIMATES OF STRENGTH OF ASSOCIATION OF INFRASTRUCTURE TERMS TO THEME PHRASE

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - EXAMPLE (NEAR-EARTH SPACE STUDY)**
- RELATION AMONG TECHNICAL THEMES AND INFRASTRUCTURE (***REMOTE SENSING***)
  - AUTHORS (CRACKNELL-AP, VARTSOS-CA, KONDRATEV-KY, GUSHIN-GA, ZAKHAROV-MY, LUPYAN-EA)
  - JOURNALS (PHOTOGRAMMATIC ENGINEERING, JOURNAL OF PHOTOGRAMMETRY, IGARRSS, IEEE TRANSACTIONS [ON GEOSCIENCE AND REMOTE SENSING])
  - INSTITUTIONS (UNIV-DUNDEE, INST MARINE HYDROPHYS SEVASTAPOL UKRAINE, UNIV DELAWARE, BOSTON UNIV, UNIV OF HAMBURG)

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - PROCESS**
- TAXONOMIES
  - TOP-DOWN
    - VISUAL INSPECTION OF THEMES
  - -BOTTOM-UP
    - SELECT MANY THEMES
    - GROUP INTO CATEGORIES USING CLUSTERING

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - EXAMPLE (SPACE STUDY)**
- TOP-DOWN SPACE TAXONOMY - SCI - PHRASE FREQUENCY BASED

- *SPACE PLATFORM (E.G., SATELLITE, SPACECRAFT)
- *SATELLITE FUNCTION (E.G., MAPPING, NAVIGATION)
- *SATELLITE TYPE (E.G., GEOSAT, LANDSAT)
- *MEASURING INSTRUMENT (E.G., RADIOMETER, MICROWAVE IMAGER)
- *REGION EXAMINED (E.G., SEA, BOUNDARY LAYER)
- *LOCATION EXAMINED (E.G., NORTH ATLANTIC, SOUTHERN HEMISPHERE)
- *VARIABLE MEASURED (E.G., TEMPERATURE, SOIL MOISTURE)
- *VARIABLE DERIVED (E.G., RADIATION BUDGET, GENERAL CIRCULATION)
- *ANALYTICAL TOOL (E.G., DATA PROCESSING, MATHEMATICAL MODELS)
- *PRODUCTS (E.G., TIME SERIES, SEA ICE MAPS)
- *SPACE ENVIRONMENT (E.G., SOLAR WIND, MAGNETIC FIELD)

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - EXAMPLE (SPACE STUDY)**
- TOP-DOWN SPACE TAXONOMY - EC - PHRASE FREQUENCY BASED
- SAME AS 1A, BUT ADD:
- *SATELLITE CONFIGURATION (GEOSTATIONARY SATELLITES, TETHERED SATELLITE SYSTEM)
- *SATELLITE STATE (ATTITUDE DETERMINATION, HIGH ELEVATION ANGLE)
- *SATELLITE SUBSYSTEMS (SOLAR CELLS, ATTITUDE CONTROL SYSTEM)

# CAPABILITIES/ EXAMPLES

## COMPUTATIONAL LINGUISTICS - EXAMPLE (HYPERSONIC/ SUPERSONIC STUDY)

BOTTOM-UP HYPERSONICS/ SUPERSONICS TAXONOMY

-SCI

| THEME NAME / #OVERLAPS | 60 | 50 | 40 | 35 | 30 | 25 | 20 | 15 |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| SHOCK TUNNEL | A60 | A50 | A40 | A35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| HIGH TEMPERATURE | A60 | A50 | A40 | A35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| HYPERSONIC FLIGHT | | A50 | A40 | A35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| BOUNDARY LAYERS | | | B40 | B35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| OBLIQUE SHOCK | | | B40 | B35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| REYNOLDS NUMBER | | | B40 | B35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| TRAILING EDGE | | | B40 | B35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| SUPERSONIC AIRCRAFT | | | | B35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| PRESSURE GRADIENT | | | | B35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| SUPERSONIC SPEEDS | | | | B35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| TURBULENT BOUNDARY LAYER | | | | B35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| VISCOUS SHOCK LAYER | | | | C35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| PERFECT GAS | | | | C35 | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |
| SURFACE PRESSURE | | | | | ABC30 | ABCDE25 | ABCDEFGH20 | ABCDEFGHIJK15 |

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - PROCESS**
- GLOBAL LEVELS OF EMPHASIS
  - IDENTIFY SINGLE, ADJACENT DOUBLE, ADJACENT TRIPLE PHRASES OF INTEREST
  - DEVELOP 'TOP-DOWN' OR 'BOTTOM-UP' TAXONOMIES IN WHICH TO GROUP PHRASES, DEPENDING ON STUDY OBJECTIVES
  - 'BIN' PHRASES AND ASSOCIATED FREQUENCIES INTO TAXONOMY CATEGORIES
  - SUM FREQUENCIES OF PHRASES IN EACH CATEGORY
  - PROVIDES ESTIMATES OF LEVELS OF EMPHASIS ON GLOBAL BASIS
  - NEEDS COMPARISON WITH REQUIREMENTS/ OPPORTUNITIES FOR CONTEXT

# CAPABILITIES/ EXAMPLES

## COMPUTATIONAL LINGUISTICS - EXAMPLE - GLOBAL LEVELS OF EMPHASIS

### SCI

- **Structures**: Strength, Design/analysis, crack initiation & growth, loads & dynamics, fatigue.
- **Aeromechanics:** Aerodynamics; Design/Analysis; Performance(A/C); Drag Reduction; Wing Design; Unsteady Flow; High Lift; Wind Tunnel
- **Subsystems:** Control Systems; Neural Nets; Environmental Control Systems; Landing Gear; Subsystems (Gen.); Actuators
- **Flight Dynamics:** Stability & Control; Helicopter Rotors; Handling Qualities
- **Systems Engineering**: Fighter/Attack; Cockpit Noise; Patrol/Transport; Conceptual Design; Air Traffic Control; Airport Noise
- **Propulsion & Power**: Gas Turbine Engine; Fuels/Lubricants; Electrical Generation; Coatings; Blades/Disks; Propeller/Propfan; Electrical Power (General); Contrails
- **Avionics:** Navigation & Guidance; Decision Aids(Processing); Avionics (Gen); S/W Development; GPS; Neural Nets; Air Data; Software/Hardware(S/W)

### EC

- **Aeromechanics**: Aerodynamics, Design/analysis, Performance(A/C), Wing Design, wind tunnel, drag reduction.
- **Structures:** Design/Analysis; Loads & Dynamics; Structures(Gen.); Crack Initiation & Growth; Strength; Structural Life; Aeroelastic Effects
- **Subsystems:** Control Systems; Environmental Control Systems; Neural Nets; Landing gear; Subsystems(Gen.); Fuzzy Logic; Actuators
- **Systems Engineering:** Conceptual Design; Fighter/Attack; Patrol/Transport; Air Traffic Control; Rotorcraft; UAV/UCAV; V/STOL
- **Avionics:** GPS; navigation & Guidance; Avionics(Gen.); Communication Systems; Artificial Intelligence; INS; Software/Hardware(S/W); Decision Aids(Processing); Information Management
- **Flight Dynamics:** Stability & Control; Helicopter Rotors; Handling Qualities
- **Propulsion & Power:** Gas Turbine Engine; Engines(Gen.); Electrical Power(General); Fuels/Lubricants; Electrical Generation; Blades/Disks

# CAPABILITIES/ EXAMPLES

## COMPUTATIONAL LINGUISTICS - EXAMPLE - GLOBAL LEVELS OF EMPHASIS

### SCI

- **Materials:** Composites; Metals/Alloys; NDI/NDT; Corrosion; Adhesives; Ceramics
- **Support/Logistics**: Maintenance; Take-off & Landing; Safety (Maintenance); Platform Interface; Deicing
- **Manufacturing:** Joints; Processes; Structural(Mfg); Concurrent Engineering; Composites(Mfg.)
- **Training:** Local Simulation; Manned Flight Simulation; Types(Instruction)
- **Costing:** Life Cycle Costs; Affordability of New Systems
- **Crew Systems:** Human/Machine Interface; Decision Aids; Loss of Consciousness

### EC

- **Materials:** Composites; Metals/Alloys; NDI/NDT; Materials(Gen); Corrosion; Smart Materials
- **Support/Logistics:** Maintenance; Reliability; Take-off & Landing; Support/Logistics(Gen.); Runaways/Airfields
- **Crew Systems:** Displays; Decision Aids' Human/Machine Interface; Data/Information Fusion; Crew Worrkload; Cockpit
- **Manufacturing:** Processes; Composites(Mfg.); Concurrent Engineering; Joints
- **Costing:** Life Cycle Costs: Affordability of New Systems
- **Training:** Simulation(Gen.); Manned Flight Simulation; Instruction(Gen.); Distributed Simulation

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - BENEFITS**
- **PHRASE FREQUENCY ANALYSIS**
  - ALLOWS LEVELS OF EMPHASIS/ EFFORT IN SPECIFIC SUBCATEGORIES TO BE ESTIMATED THROUGH 'BINNING’
  - ALLOWS JUDGEMENTS OF ADEQUACY AND DEFICIENCY IN SELECTED S&T AREAS TO BE MADE ON GLOBAL BASIS
    - NEEDS COMPARISONS TO REQUIREMENTS/ OPPORTUNITIES FOR JUDGEMENT CONTEXT
  - PROVIDES COMPREHENSIVE PICTURE OF MAJOR THRUST AREAS

# CAPABILITIES/ EXAMPLES

- NO RELATIONAL INFORMATION; NOT USEFUL FOR ESTIMATING LINKAGE BETWEEN S&T AREAS

- ***USEFUL TO APPLY TO MULTIPLE DATABASE FIELDS TO GAIN DIFFERENT PERSPECTIVES; FIELDS USED FOR DIFFERENT PURPOSES***
  - KEYWORDS
  - ABSTRACTS
  - TITLES

- AIRCRAFT EXAMPLE
  - LONGEVITY AND MAINTENANCE IN KEYWORDS
  - NO PERFORMANCE IN KEYWORDS
  - NO TESTING IN KEYWORDS
  - OTHER AREAS SIMILAR (MATERIALS/ CONTROLS, ETC)

# CAPABILITIES/ EXAMPLES

- **COMPUTATIONAL LINGUISTICS - BENEFITS**
  - **PHRASE PROXIMITY ANALYSIS**
  - ACCESS COMPLEMENTARY LITERATURES WITH RELATED THEMES
    - HIGH POTENTIAL FOR INNOVATION AND DISCOVERY FROM OTHER DISCIPLINES
  - ALLOWS INFRASTRUCTURE (AUTHORS/ JOURNALS/ ORGANIZATIONS) RELATED TO SPECIFIC TECHNICAL AREAS TO BE IDENTIFIED
  - ALLOWS CLOSELY RELATED THEMES TO BE IDENTIFIED
    - POTENTIAL FOR IDENTIFYING "NEEDLE-IN-A-HAYSTACK"

# CAPABILITIES/ EXAMPLES

- ALLOWS TAXONOMIES WITH RELATIVELY INDEPENDENT CATEGORIES TO BE GENERATED USING A 'BOTTOM-UP' APPROACH
  - STARTS WITH MANY HIGH FREQUENCY THEMES
  - GROUPS RELATED THEMES INTO CATEGORIES USING PROXIMITY ANALYSIS
  - **SEE JASIS PAPER (15 APRIL 1999**) FOR DETAILED EXAMPLE OF TAXONOMY GENERATION
  - PRESENTLY DEVELOPING MORE AUTOMATED CLUSTERING APPROACH USING CO-OCCURRENCE MATRICES

- USEFUL FOR ESTIMATING LEVELS OF EMPHASIS CLOSELY ASSOCIATED WITH THE THEME

# OUTLINE

- DEFINITIONS/ GOALS
- CAPABILITIES/ EXAMPLES
- **CROSSOVER SCIENCE**
  - BACKGROUND
  - CONCEPT
  - PROPOSAL
- DEFICIENCIES
- NEXT STEPS
- SUMMARY

# CROSSOVER SCIENCE

- **CONCEPT**
  - LINK MULTIPLE DISJOINT LITERATURES THROUGH INTERMEDIATE LITERATURES
  - A--->B; B--->C; A===>C
  - DISCOVERY FROM REMOTE LITERATURES COULD NOT HAVE BEEN OBTAINED FROM PRIME LITERATURE

# CROSSOVER SCIENCE

- **BACKGROUND**
  - SWANSON PUBLISHED APPLICATIONS IN MID-1980S (DESCRIBE)
  - FOCUSED ON MEDICAL LITERATURE AND MEDLINE DATA BASE
  - OUR GROUP PUBLISHED CONCEPT PAPER IN 1999, IN TECHNOVATION
  - PROPOSED DEMONSTRATION ON BIOLOGICAL WARFARE AGENT PREDICTION

# CROSSOVER SCIENCE

- **PROPOSAL** (DISCOVERY FROM LITERATURE COMPONENT)

- DEFINE TARGET LITERATURE THAT DESCRIBES WHAT WE KNOW
- USING COMPUTATIONAL LINGUISTICS, IDENTIFY CHARACTERISTIC FEATURES OF THAT LITERATURE
- GENERATE LITERATURES CENTERED AROUND THE CHARACTERISTIC FEATURES (E.G., VIRULENCE, TRANSMISSIBILITY)
- FORCE EACH LITERATURES TO BE DISJOINT FROM TARGET LITERATURE BY ELIMINATING INTERSECTION
- USING COMPUTATIONAL LINGUISTICS, IDENTIFY CANDIDATE VIRUSES IN EACH CHARACTERISTIC FEATURE LITERATURE
- REMOVE ALL COMMON PHRASES BETWEEN TARGET LITERATURE AND EACH CHARACTERISTIC FEATURE LITERATURE
- COMBINE LISTS OF CANDIDATE VIRUSES FROM EACH CHARACTERISTIC FEATURE LITERATURE INTO ONE CANDIDATE VIRUS LIST
- ASSIGN SCORES TO CANDIDATE VIRUSES, BASED ON NUMBER OF TIMES THEY APPEAR IN LIST, VALUE OF NUMERICAL INDICATORS FROM COMPUTATIONAL LINGUISTICS, AND PRIORITY WEIGHTING ASSIGNED TO IMPORTANCE OF EACH CHARACTERISTIC FEATURE.
- RECOMMEND HIGHEST RANKED VIRUSES.

# CROSSOVER SCIENCE

- **DIFFERENCES WITH SWANSON APPROACH**

- 1) HE FOCUSES ON TITLES; WE FOCUS ON ABSTRACTS, BUT COULD JUST AS EASILY USE FULL TEXT IF AVAILABLE
- 2) HE FOCUSES ON MEDLINE; WE CAN USE OTHER DATABASES, MOST NOTABLY SCI, IF WARRANTED BY THE CHARACTERISTIC FEATURES IDENTIFIED FROM THE COMPUTATIONAL LINGUISTICS OF THE TARGET LITERATURE
- 3) HE USES MESH IDENTIFIERS; WE USE DIRECT TEXT PHRASES
- 4) HE USES QUERY TERMS AB INITIO; WE USE AN ITERATIVE LITERATURE BASED QUERY DEVELOPMENT
- 5) HE DEFINES THE CHARACTERISTIC FEATURES AB INITIO; WE USE COMPUTATIONAL LINGUISTICS ON EXPERT-GENERATED RELEVANT LITERATURE TO DEFINE CHARACTERISTIC FEATURES
- 6) THERE IS ALSO A DIFFERENCE IN HOW WE EMPLOY COMPUTATIONAL LINGUISTICS
- 7) HE HAS PUBLISHED RESULTS OF HIS DISCOVERY TECHNIQUE IN THE LITERATURE, WHILE WE HAVE PUBLISHED ONLY RESULTS OF OUR STANDARD TEXT MINING TECHNIQUE.

# OUTLINE

- DEFINITIONS/ GOALS
- CAPABILITIES/ EXAMPLES
- CROSSOVER SCIENCE
  - BACKGROUND
  - CONCEPT
  - PROPOSAL
- **DEFICIENCIES**
- NEXT STEPS
- SUMMARY

# DEFICIENCIES

- MOTIVATION
- PERSONNEL
- INFORMATION EXTRACTION
- DATABASE AVAILABILITY
- STRATEGIC MANAGEMENT INTEGRATION

# DEFICIENCIES
# **MOTIVATION**

- LACK OF MOTIVATION TO DEVELOP/ DEMONSTRATE/ USE S&T TEXT MINING
  - LACK OF DEVELOPMENT SUPPORT
  - LACK OF INDIVIDUAL USER SUPPORT
  - LACK OF MANAGEMENT USE

# DEFICIENCIES
# **PERSONNEL**

- FEW PEOPLE INVOLVED IN DEVELOPING TM
  - REQUIRES TEAM OF
    - DISCIPLINE TECHNICAL EXPERTS
    - EXTRA-DISCIPLINE TECHNICAL EXPERTS
    - INFORMATION TECHNOLOGISTS
  - LITERATURE-BASED DISCOVERY
    - ONE GROUP PUBLISHING
    - PERHAPS THREE GROUPS INVOLVED

# DEFICIENCIES
# **INFORMATION EXTRACTION**

- SEMI-AUTOMATED PHRASE EXTRACTION ALGORITHMS INCOMPLETE
- EXTENSIVE MANUAL CLEANUP REQUIRED
- POOR PHRASE GENERATION LEADS TO:
  - LOST QUERY TERMS FOR INFORMATION RETRIEVAL
  - LOST CONCEPTS FOR LITERATURE-BASED DISCOVERY
  - INCOMPLETE TAXONOMIES FOR DISCIPLINE CLASSIFICATION
  - INCORRECT CONCEPT CLUSTERING

# DEFICIENCIES
# **CLUSTERING**

- LITERATURE FOCUS ON DOCUMENT CLUSTERING
- CONCEPT CLUSTERING CAN PROVIDE INSIGHTS
- CLUSTERING QUALITY DEPENDS ON:
  - AGGLOMORATION TECHNIQUES
  - ASSOCIATION METRICS
  - QUALITY OF PHRASES
  - COMPLETENESS OF PHRASES
  - THRESHOLD CRITERIA
  - NUMBER OF PHRASES
- SUBSTANTIAL TIME AND EFFORT REQUIRED
  - CLEANUP/ INTERPRETATION

# DEFICIENCIES
# **DATABASE**

- SMALL FRACTION OF S&T PERFORMED AVAILABLE TO TEXT ANALYST
  - SMALL FRACTION OF S&T DOCUMENTED
  - SMALL FRACTION OF DOCUMENTATION INCLUDED IN DATABASES
  - MODEST FRACTION OF DATABASES ACCESSIBLE
    - RELATIVELY HIGH COST
    - NOT WELL ADVERTISED
    - NON-STANDARD INTERFACES
    - SEARCH ENGINES UNFRIENDLY
  - POOR INFORMATION RETRIEVAL TECHNIQUES USED

# DEFICIENCIES

## STRATEGIC MANAGEMENT INTEGRATION

- TEXT MINING CONDUCTED IN ISOLATION FROM STRATEGIC MANAGEMENT
  - IDEALLY
    - OBJECTIVES -> METRICS -> DATA
  - PRESENTLY
    - DATA -> METRICS -> OBJECTIVES
- PART OF LARGER PROBLEM WITH ALL MANAGEMENT DECISION AIDS

# OUTLINE

- DEFINITIONS/ GOALS
- CAPABILITIES/ EXAMPLES
- CROSSOVER SCIENCE
  - BACKGROUND
  - CONCEPT
  - PROPOSAL
- DEFICIENCIES
- **NEXT STEPS**
- SUMMARY

# NEXT STEPS

- TECHNOLOGY UPGRADES
  - AUTOMATE MARGINAL UTILITY
    - GENERATE OPTIMAL QUERIES
  - ADD CLUSTERING
    - SHORTEN QUERY DEVELOPMENT
    - IMPROVE TAXONOMY DEVELOPMENT
    - IDENTIFY THEME LINKAGES FOR DISCOVERY
  - ADD FUZZY LOGIC
    - IMPROVED BIBLIOMETRICS
  - ADD CO-OCCURRENCE
    - ELIMINATE EXTRA PLATFORM
    - IMPROVE THEME LINKAGES

# NEXT STEPS

- TEXT MINING STUDIES USING UPGRADED TECHNOLOGY
  - INFORMATION RETRIEVAL
  - BIBLIOMETRICS
  - PHRASE FREQUENCY ANALYSIS
  - PHRASE PROXIMITY ANALYSIS

# NEXT STEPS

- CROSSOVER SCIENCE
  - USE UPGRADED TECHNOLOGY
  - USE NEW CONCEPTS/ CLUSTERING
  - BIOWARFARE AGENT PREDICTION
  - (PROPOSAL-HAVE TEAM)
- CITATION MINING
  - IDENTIFY DOCUMENTED USERS
  - IDENTIFY IMPACTS OF RESEARCH

# OUTLINE

- DEFINITIONS/ GOALS
- CAPABILITIES/ EXAMPLES
- CROSSOVER SCIENCE
  - BACKGROUND
  - CONCEPT
  - PROPOSAL
- DEFICIENCIES
- NEXT STEPS
- **SUMMARY**

# SUMMARY

- GLOBAL TECHNOLOGY WATCH CRITICAL
- TEXT MINING CAN IDENTIFY RELEVANT LITERATURE/ EXTRACT INFORMATION
- NEED TO OVERCOME BARRIERS IN:
  - LACK OF MOTIVATION
  - LACK OF PERSONNEL
  - INFORMATION EXTRACTION TECHNIQUES
  - DATABASE AVAILABILITY
  - INTEGRATION WITH STRATEGIC MANAGEMENT
- OUR GROUP'S FOCUS
  - UPGRADE SOFTWARE TECHNOLOGY
  - APPLY TO **OUR** STANDARD TEXT MINING
  - EXPAND CROSSOVER SCIENCE
  - DEMONSTRATE CITATION MINING

# TRACK RECORD

- DEVELOPED FULL TEXT CO-WORD TEXT MINING FOR S&T EVALUATION
  - PREVIOUS EFFORTS USED KEY WORDS ONLY
- PUBLICATIONS
  - 16 PAPERS IN PEER REVIEWED JOURNALS
  - 9 PAPERS IN PEER REVIEWED CONF. PROCEED.
  - 1 BOOK CHAPTER
  - 2 PAPERS ON WEB SITES
  - 4 PAPERS SUBMITTED TO JOURNALS
  - 10 PAPERS TO BE SUBMITTED TO JOURNALS
- JOURNALS
  - JASIS, IPM, JIS (INF TECH)
  - CHEMICAL REVIEWS, JOURNAL OF AIRCRAFT, ANALYTICAL CHEMISTRY (NON-INF TECH)

# TRACK RECORD

- **TOAS/ IFO**
  - PATENTED SOFTWARE LENT TO TOAS DEVELOPMENT GROUP IN MID-1990S
  - ONR TEXT MINING PAPERS CITED 14 TIMES BY TOAS DEVELOPERS IN PUBLISHED LITERATURE
  - CORRESPONDENCES STIMULATED IFO ENTRY INTO TEXT MINING
- **ONR/ IFO**
  - PILOT PROGRAM PROPOSAL IN DECEMBER 1997 STIMULATED ONR ENTRY INTO TEXT MINING
  - ACCELERATED IFO PROGRESS IN TM